

DOCUMENT RESUME

ED 285 169

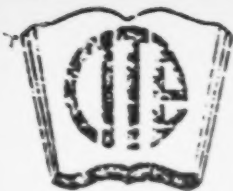
CS 210 671

AUTHOR Butler, Sydney J.
TITLE A Survey of the Testing of English/Language Arts in British Columbia, 1984-85.
INSTITUTION Committee to Improve the Teaching of English (British Columbia).
PUB DATE 20 Jan 86
NOTE 34p.; Interim report of the Committee to Improve the Teaching of English Commission on the Testing of English/Language Arts in British Columbia.
PUB TYPE Reports - Research/Technical (143) -- Tests/Evaluation Instruments (160)
EDRS PRICE MF01/PC02 Plus Postage.
DESCRIPTORS Achievement Tests; Educational Research; Elementary Secondary Education; *English Instruction; Foreign Countries; Higher Education; *Language Arts; School Surveys; Standardized Tests; Student Evaluation; *Testing Programs; *Test Use
IDENTIFIERS British Columbia; *Canadian Achievement Tests

ABSTRACT

A survey was conducted to ascertain the amount and type of external testing (defined as any test introduced from outside the classroom) of English/language arts being administered throughout the province of British Columbia (Canada) during the 1985-1986 school year. Teachers and administrators were also given the opportunity to comment on the testing programs. Respondents, representing 961 schools and colleges (out of 1567 in the survey--a 61.3% return rate) as well as local teachers' associations, completed a questionnaire on testing. Results indicated a wide variety in the use of tests. A considerable proportion of schools (9.1%) indicated an avoidance of all externally imposed tests, while a similar percentage (9.8%) reported using only the tests mandated by the Ministry of Education. The remainder was divided between tests imposed by either the local district or school administrators. Results did not reveal the content tested in the locally developed tests. Findings also showed the widespread use of standardized tests, such as the Canadian Test of Basic Skills (a Canadian adaptation of the Iowa Tests of Basic Skills) or the Gates-MacGinitie Reading Test. (References, tables of data, and a sample questionnaire are attached.) (NKA)

* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *



COMMITTEE TO IMPROVE THE TEACHING OF ENGLISH

Chairperson: (Dr.) Jerry Zaslove, English Department, SFU

Secretary: John Church, Professional Development, SCIS

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES INFORMATION
CENTER (ERIC)

X This document has been reproduced as
received from the person or organization
originating it.

□ Minor changes have been made to improve
reproduction quality.

• Points of view or opinions stated in this docu-
ment do not necessarily represent official
OERI position or policy.

ED285169

A SURVEY OF THE TESTING OF ENGLISH/LANGUAGE ARTS IN B.C., 1984-85

The Interim Report of the CITE Commission on the Testing
of English/Language Arts in B.C., as presented to the
Committee on January 20, 1986, and as discussed and analyzed
by Dr. Ted Aoki in a further meeting on March 7, 1986.

"PERMISSION TO REPRODUCE THIS
MATERIAL HAS BEEN GRANTED BY

Sydney J. Butler

TO THE EDUCATIONAL RESOURCES
INFORMATION CENTER (ERIC)."

2

BEST COPY AVAILABLE

CS 210 671

THE COMMISSION ON TESTING OF ENGLISH/LANGUAGE ARTS

Berenice Wood
B. C. Teachers' Federation

Sydney Butler
Language Education Dept,
University of B. C.

Brian Marrs
College-Institute
Educators'
Association

A. SURVEY OF THE TESTING OF ENGLISH/LANGUAGE ARTS

IN BRITISH COLUMBIA

1984-85

Dr. Sydney J. Butler
Language Education Department
University of British Columbia
Vancouver, B.C. V6T 1Z5

THE TESTING OF ENGLISH/LANGUAGE ARTS IN B.C.

BACKGROUND TO THE STUDY

The Committee to Improve the Teaching of English is a joint body of representatives from elementary and secondary schools, colleges, universities, and the Ministry of Education in British Columbia. CITE meets several times each year to discuss current issues affecting the teaching of English or language arts at all levels of education.

One area of concern which came to prominence with the re-introduction of the provincial grade 12 examinations in 1984 is the amount of testing of English throughout the school and college systems. In addition to the tests, examinations, or assessments prescribed or recommended by the Ministry of Education, individual reports suggested that many schools or school districts were also using locally-developed instruments or commercial, standardized tests to assess the students' abilities in the use of language.

For example, at the grade 12 level students may be required to write, at the discretion of their teacher, the Classroom Achievement Test in Written Composition developed by the Ministry of Education. In addition, all students wishing to graduate, except for those in the "Minimum Essentials" category, must take the provincial English 12 examination, which counts for one half of the final course mark. Even such courses as Communications 11 and Communications 12 may soon be destined to culminate in a provincial examination. Many of the students intending to continue into post-secondary education in British Columbia must also take the English Placement Test, or a similar local assessment test, while those who are interested in American colleges prepare for the SAT tests.

Such a plethora of tests added to the teacher-made tests which are a normal part of any course may suggest some overkill in evaluation, especially as there is considerable overlap in the goals and objectives of those tests or sections which deal with the students' abilities in written expression.

On the other hand, CITE was concerned that none of these tests seemed to take into account the current state of knowledge about the assessment of written composition, especially the unreliability of using only one writing sample from each student, the variability inherent in the choice of topic, and the difficulties of maintaining reliable, consistent ratings of student writing.

CITE also expressed concern about the effects of such testing practices on the school system as a whole. For example, it may be shown that such tests measure only a narrow band of the English curriculum, neglecting completely the students' oral development. Yet teachers almost inevitably will be inclined to "teach to the test." In this way tests may bias the English curriculum in the direction of writing only in the particular modes required by the tests, and reading only for the lower levels of comprehension that can be tested by multiple-choice, objective tests.

Another area of the Committee's concern about the use of standardized tests was whether such tests recognize the particular ethnic and cultural mix

of British Columbian students. The norms established by reference to a sample population of North American students, or even a Canadian population, may not be valid for students in this province. In fact, the Ministry of Education's resource book *Selection and Use of Standardized Tests* (Victoria, 1979) cautions teachers and administrators to examine the adequacy of the norm-referencing by which a test was standardized. Unless the test items and scoring scales are adapted for local norms, the reliability and validity of the test result may be suspect.

Accordingly, CITE called for the establishment of a task force or commission to gather information on all large-scale testing of English in British Columbia, including an evaluation of the test instruments and the procedures used to select these instruments, and to compare the testing objectives and practices of British Columbia with those of other Canadian or foreign programs of large-scale assessment.

Consequently the CITE Commission on the Testing of English/Language Arts was established in March, 1985, with the following mandate:

1. Determine what external tests of English or language arts, including reading, are currently being used in B.C. schools and colleges.
2. Determine how such tests are chosen and for what purposes.
3. Assess the validity and reliability of such tests, particularly to examine the relationship of such tests to the English and language arts curricula.
4. Involve parents, teachers, and administrators in a discussion of the effects of such testing programs.

THE SURVEY OF EXTERNAL TESTS USED IN B.C., 1984-85

Through the cooperation of the B.C. Teachers' Federation, the College-Institute Educators' Association, and the Language Education Department of the University of B.C., the Commission prepared a questionnaire (see Appendix A) which was mailed to the 1165 elementary schools, 281 secondary schools, 103 combined elementary/secondary schools, and 18 colleges in British Columbia, together with letters to the principal and B.C.T.F. representative in each school, to the superintendents of the 75 school districts, and to the presidents of the local teachers' associations, explaining the purpose of the survey and asking for their cooperation.

For the Survey, the Commission used the term "external test" to mean any test introduced from outside the classroom, whether by the choice of the teacher or by an administrator. External tests could thus be tests prepared by the Ministry, ERIBC, district or school personnel, or those purchased from commercial publishers. The Survey would not therefore be concerned with the tests which teachers prepared as part of their normal evaluation procedures in their classrooms. The Commission recognized that teachers regularly use tests and other instruments to diagnose their students' abilities, to assess their achievements, and to judge the effectiveness of their programs and curricula. Such a program of evaluation is an essential component of teaching, but the mandate of the Commission is to review the tests which are superimposed on the regular program of classroom evaluation.

RESULTS OF THE SURVEY

From the 1567 schools and colleges in the Survey the Commission received 961 responses (61.3 per cent). Eighty-eight of these responses (9.1 per cent) showed a null return, although comments on several of these responses indicated that the respondents did not take the questionnaire to mean to include those government exams which are a part of the normal school routine. One respondent struck a plaintive note, as if the respondent felt that the school was missing out on something, with the comment: "We don't seem involved." Another respondent noted that although there was nothing to report for the current school year, there had been such tests imposed in previous years, but that this year was a "hiatus in the district testing program." Several of the null returns proclaimed the school's freedom from external testing, a typical comment which strikes a note of independent pride being: "We have had no such tests imposed. We do our own evaluation within our department and as individual teachers."

Although the primary purpose of the Survey was to ascertain the amount and type of external testing of English/language arts being conducted throughout the province during the 1985-86 school year, teachers and administrators also had the opportunity to comment about the testing programs. While most of these comments were brief explanatory notes written during the hectic bustle of a school day, several respondents took the opportunity to air more extensive views about the effects of testing programs on their students. The most comprehensive and analytical comments were critical of testing programs as applied at either end of the school

system. A grade 3 teacher complained about the psychological damage that a standardized language test was having on the children, while a college instructor gave a lengthy rationale to explain why that institution had stopped using a common composition test for all first year students. One secondary English teacher complained about the inefficiency of the English Placement Test, while another gave a full page of specific charges against the provincial English 12 and Literature 12 examinations.

The grade 3 teacher explained that she had used a standardized reading test herself for many years, but used the results merely as a back-up to her own assessments of the students' reading abilities. She noted that during the tests "the low-ability, behavior-problem, or underachieving bright children rarely perform well . . . they rush through and fill in any answer just to get finished. They haven't learned 'appropriate test-taking behavior.'" She warns of the danger of using unreliable test scores obtained under such conditions as indicators of the students' abilities. Moreover, when her class had to take an externally-imposed battery of tests over a week in June she found that the effects were more immediately deleterious. "One of the children . . . actually threw up on the first day of the testing. He was so worked up and just couldn't handle it." Another child she noticed "didn't seem to have the desire to even complete more than 10 questions out of 30 in each section." Consequently, when at parent-teacher conferences she showed the children's results plotted against the "grade level average," she had to explain why some of her slower children showed up as being one grade or more below the norm. Perhaps it was fortunate that this teacher was at least aware of the unreliability of the test results, and was also able to explain the built-in Standard Error of Measurement factor of this particular test, which in her case was compounded by the frustration caused by the testing situation.

At the college level, an instructor explained that his or her department had abandoned the external test, because the teachers had found "it was basically a waste of time in that the results almost always merely confirmed the class instructor's assessment of the students' abilities." This respondent also noted that the test served "no useful educational purpose Other than the score, students received little or no feedback."

A secondary English teacher also noted that the results of the English Placement Test seemed only to confirm the assessments already made by the students' grade 12 teacher. He or she complained that the test was not only a waste of the students' time, but also it was highly "resented by students and teachers alike," especially when the students have to pay for the doubtful privilege of writing the test "with minimal supervision at another institution which made little attempt to create a test atmosphere." This respondent commended S.F.U. and the "few enlightened institutions" which ignore the E.P.T. entirely.

The criticism of the English 12 and Literature 12 examinations was directed, not at the reliability or validity of the tests themselves, but rather at the effects that these examinations were having on the teaching of these courses in the school. The respondent noted that teachers were already giving intensified instruction in short essays, poetry analysis (even for vocational students), grammar and usage, with the consequence that there was reduced emphasis on oral activities (speeches, debates, etc.), elimination from

the syllabus of extended studies of a novel and a play, and a general reduction in enrichment activities and even the "job-search skills unit for vocational students." He commented on the "strong variation in student motivation" depending on whether a particular activity is "on the exam or not."

In the Literature 12 class the examination was having the effect of increased emphasis on "objective memorization" and a reduction in written assignments, research papers and essays, with the elimination of all "supplemental and enrichment activities, films, and readings." The examination "had reduced to nil extended classroom analysis and discussion of materials." With the reduction in the quality of the curriculum the respondent noted that there had been a fifty per cent reduction in the number of students electing to take this course.

THE SCOPE OF TESTING PROGRAMS

Because the Survey was designed to cast a wide net in order to compile a general view of language testing throughout the Province, the results show a great variety in the type, purpose, and scope of the reported testing programs. While the term "language test" would obviously net the big fish of such test batteries as the *Metropolitan Achievement Test*, *The Canadian Achievement Tests*, or *The Canadian Test of Basic Skills*, all of which contain a separate section under the rubric of "Language," it was also possible to capture more specialized aspects of language testing. Consequently, while some respondents to the Survey reported on tests designed to measure the achievement of students in various secondary English courses, others reported on tests designed to measure such specific abilities as vocabulary, written expression, spelling, syntax, auditory discrimination, or listening comprehension.

Reading is another aspect of language ability which is covered in multi-test batteries, usually under such headings as phonic skills, word recognition and reading comprehension. In addition, the Survey also showed that a number of more specialized test instruments were being used. Several teachers reported the use of criterion-referenced tests provided by publishers to show a student's mastery of reading skills at a particular level of a basal reading program. Others reported the use of special tests to measure reading readiness, word recognition, basic sight words, or industrial reading. Moreover, most tests of "basic concepts," "psycholinguistic abilities," "school abilities," or "critical thinking" contain an element of reading comprehension.

The Survey also showed differences between programs implemented at a school district level involving students in specified grades at a number of schools, and programs chosen by school staff or administrators for testing all or some of the students in that particular school. In many cases school administrators used the resources of the district administration in helping them to select specific test instruments. In other cases an individual teacher was able to use the district resources to select a test for a particular class or group of students. However, the Survey failed to distinguish between district programs which used sampling techniques to select certain classes or schools as representative samples of a whole population, and those programs

which blanketed all of the students at specified grade levels within the district.

DISTRICT TESTING PROGRAMS

The Survey shows a considerable amount of testing being conducted at the district level. One district was reported to have conducted a test of reading which was compulsory for all students in grades 3-7. The test was chosen or developed by district administrators, marked by machine-scoring, and the results tabulated and distributed by computer printout. Because the results listed the scores of individual pupils, the purpose of this testing program was not clearly defined in the responses to the Survey.

While some regarded this test as an assessment of reading programs, many respondents noted that the purpose of the test was to assess and diagnose the abilities of individual students. It seems that the District staff may use these results to compare the performances of various schools and classes in the District, while many teachers might be following the practice of one respondent who noted that the tests "are used by the individual teacher to assess and diagnose pupils' achievement and placement." On the other hand, another respondent noted that the tests are specifically designed for program assessment, and commented that "teachers sometimes erroneously use them for pupil assessment."

This may be a case of the District trying to kill two birds with one stone. If the aim were simply to compare the reading performances across various schools and classrooms, then it would not be necessary to test every student at the designated grade levels. Instead, a representative random sample of students would provide the desired information. Yet, if the tests were regarded as measurements of individual pupil achievement then it would be necessary to check the validity of the test instrument against the aims and objectives of the whole reading program in order to see whether the test would be likely to give a fair assessment of an individual student, or whether the teacher's on-going evaluation procedures would give a more valid assessment, including perhaps an assessment of the student's attitudes to reading generally, and the quality and variety of reading achieved both within and outside the classroom. Such important dimensions of the provincial curriculum for reading are ignored by standardized, objective tests.

In another district program, reported to be measuring reading ability, the teachers and school administrators participated in the choice of the test, which was marked by the teachers themselves. In some schools the test was made compulsory for all students in grades 1-7, while other schools made it an option for the teachers to use, mainly in grades 4-7.

Reports of the purpose of these tests showed considerable disparity. Some respondents perceived the test as an assessment of the school's reading program; others apparently used the test results for pupil diagnosis and assessment, with over a half of the respondents reporting that the results were used for pupil placement, perhaps in streamed classes, or more likely in ability groups within a class.

District tests are also being used to assess writing abilities. One district reports on a compulsory testing program for all grade 10 students, measuring the students' abilities in writing in the narrative mode in the fall, and in the expository mode in the spring. The papers are marked by teachers who are given in-service time for the task, and the results compiled for schools, classes, and individual students, in order that students can be placed in appropriate courses in grade 11.

Another school district is reported to have developed a test of grade 5 writing with the participation of a task force of grade 5 teachers. This test was administered to all grade 5 students in the district and the papers marked by teams of teachers during the summer. At the district level the results are used for program assessment, but as the results of individual scores are made available to teachers, parents, and students, this test may also be used for pupil assessment, and according to one report, pupil placement.

Another district has developed its own writing diagnostic test which is compulsory for all grade 8, 9, and 10 students. The fact that this test is reported to be a machine-scored objective test suggests that it is not a direct test of students' writing in that an objective test could not demand that the students actually produce any of their own writing. Such objective tests of writing are usually more concerned with the editing skills needed to detect and correct errors in artificially-constructed sample sentences or paragraphs, rather than with the students' ability to express ideas in writing. Nevertheless, the results are compiled for schools and classes, while individual results are made available to teachers, parents, and students.

The disadvantage of administering a diagnostic test late in the school year was pointed out by one respondent who noted that when the papers were marked during the summer, nothing was done with the results for the grade 9 and 10 students. The complaints of the teachers had led to a change of policy so that grade 8 students are to be tested in September and grade 10 in January, enabling teachers to mark the tests and include the individual results in the students' report cards. The respondent commented that in this way "teachers would have the results in the same year that they are teaching and thus be able to incorporate the outcomes in their teaching strategies."

TESTS OF ENGLISH COURSES

Eight reports indicated tests that were used not simply to measure reading or writing ability, but as summative assessments of the students' learning in an English course. In one district, teachers involved in the grade 8 English course cooperated in compiling a common final exam for all the grade 8 students. In another case the district staff provided the test instrument which was marked by the teachers concerned to give the students' year-end assessment. Another district provides a common, compulsory examination for all grade 9 students, the results being used for both program evaluation and student assessment. In both of these cases there is no indication whether the examination counts for the whole of the student's final

grade, or whether, as with the English 12 examination, the student's course work counts for a half of the final mark, the remainder coming from in-term tests and assignments. On the other hand, one of the district examinations for grade 10 counts for only twenty per cent of the student's final mark. In another case only those students with less than a C+ average in course work are required to take the district examination. Similarly, the teachers at another school have compiled a common English 11 examination which is marked by the teachers to provide student assessments.

While these examinations cannot be piloted and standardized in the technical sense, they may however provide some common standards of achievement for the students of one grade level in a particular school or district. To the extent that they are compiled by teachers or district staff who are much closer to the classroom syllabus, they may be considerably more representative of the curriculum of the course being evaluated. Nevertheless, as with the provincial English 12 examination, these tests are also strictly limited in what aspects of the curriculum they can cover.

Obviously, such paper-and-pencil tests cannot cover the oral dimensions of the curriculum, but it is possible that they may focus on works of literature which are actually taught in the courses instead of being limited to short, disembodied pieces of prose or poetry, deliberately chosen from outside the scope of the normal curriculum materials, as happens in the provincial English 12 examination.

If, as in the English 12 examination, these cross-grade examinations focus on reading skills, knowledge of standard usages in editing, and the recall of objective facts about literature, then it is to be hoped that a student's mark for the course might recognize achievement in the appreciation and interpretation of literature, in writing in a variety of modes for different audiences, in ability to use language to express thoughts and feelings, in skill in speaking clearly and effectively, as well as the students' critical understanding of the use of language in communications media.

The danger with these, as with all tests originating from outside the classroom, is that the tail might wag the dog. Where the reputation of a school, a class, or a teacher, or the student's academic progress depends upon results achieved in competition with students from other classes, schools, or districts, then there is considerable pressure to put more emphasis on the teaching of those facets of English which are amenable to large-scale testing, with the consequent neglect of other aspects of English, even though these may be just as important to the growth and development of the student. When external tests are used to assess a student's progress and achievement in a course, resulting in a grade that may have a considerable impact on the student's future, then it is essential that the content of the test should match the content of the course, giving due attention to the texts and literature studied during the term or semester, and fully reflecting in its weighting of test questions the particular emphasis of the course.

COMMERCIALLY PUBLISHED STANDARDIZED TESTS

With tests that are bought off the publisher's shelf there can be no expectation that the skills and knowledge covered by the test items will match the curriculum. For the most comprehensive test the publisher can merely attempt to deal with those concepts and language conventions that appear to predominate in many textbooks or curriculum guides across a nation or a continent. Consequently the emphasis in these tests is on the more mechanical aspects of written language, especially the recognition and identification of errors in spelling, punctuation, capitalization, vocabulary, and grammar or usage. In order to standardize the scores, it is essential that such tests be administered under carefully controlled conditions, and with answers that are susceptible to statistical analysis. This type of test must necessarily be considered an *indirect* test of a student's language ability, in that the student is responding to language samples produced by the test-maker. An objective test cannot be a direct test of the student's language in which the student would actually have to produce written language, demonstrating control of the skills of spelling, punctuation, usage, diction, etc., in the language created during the course of the test.

Those tests which deal with reading frequently test the sub-skills of word recognition, phonic and structural analysis, as well as the comprehension of sentences and paragraphs. It should be noted, however, that the very nature of a standardized, objective-type test demands that answers must follow a fixed format, usually as multiple choice. Comprehension questions, therefore, must necessarily be convergent, requiring a single possible answer, rather than divergent questions which allow for individual responses. Hence, comprehension is tested mainly at the literal and inferential levels, rather than the evaluative and appreciative levels which the curriculum guide suggests should also be goals for teaching.

The responses to the Survey brought 712 reports of commercial tests being used in classrooms to measure reading, spelling, syntax, word analysis, auditory discrimination, etc. Although the returns showed the use of 42 different tests, it is apparent from Appendix B that some tests are used quite frequently, while others are chosen for such specific purposes as the diagnosis and assessment of students entering school for the first time, or for those who might need placement in special education or ESL classes. The most frequently mentioned tests are:

The Canadian Test of Basic Skills (264)
The Gates MacGinitie Reading Tests (199)
The Canadian Achievement Tests (70)
The Stanford Diagnostic Test (48)
The Nelson-Denny Reading Test (29)
The Stanford Achievement Test (23)

The questionnaires show for each reported use of these tests whether it was made compulsory for all students or not, the purpose of the test, who was responsible for choosing the test, how the test was marked, how

the results were recorded and distributed, and how much class time was involved in taking the tests. In addition, respondents were invited to make comments on the back of the questionnaire sheet.

Obviously, when the results of such tests are to be used to evaluate the effectiveness of particular programs, textbooks, or teaching styles, then it is essential that the content validity of a test is adequate for that particular district or school. The authors of the *Metropolitan Achievement Test*, for example, recognizing that no standardized commercial test can be universally valid, recommend that each school staff should consider whether their test has content validity for their curriculum. Similarly, the B.C. Ministry of Education in *Selection and Use of Standardized Tests* states that "it thus becomes an important responsibility of the test user to determine the degree to which each test's content reflects locally held curriculum emphases" (p.8).

THE CANADIAN TEST OF BASIC SKILLS

Because the *Canadian Test of Basic Skills* appears to be used widely throughout the province it is worth considering the content of this test battery. Altogether *CTBS* tests vocabulary, reading comprehension, language skills, work-study skills, and mathematics skills at eight different levels of ability. As the Survey is concerned only with the testing of language skills, it is not clear how many of the responding schools and districts administered the complete battery, requiring on the average about four hours of class time. The Survey shows that the language sections (i.e. vocabulary, reading comprehension, and the language skills of spelling, capitalization, punctuation and usage) were administered mainly across grades 1-7, with some reports of the tests being used at the junior and senior high levels.

Ninety-two per cent of the responses for *CTBS* stated that the test was compulsory for all students at the designated levels, although there are comments to the effect that some ESL or slow students were excused from taking the test. In Table 1 (Appendix C), which shows the Reported Purposes for the use of *CTBS*, it is clear that respondents regard the test mainly as an assessment of individual students, only secondarily for the purposes of comparing school or class performance, with very little consideration that the test might help to evaluate language programs. On the other hand, Table 3 (Appendix E) shows that generally teachers seemed to have little input into the choice of the test. Most (73.2 per cent) of the respondents stated that the test was selected by school or district administrators, and that the marking was done by machine-scoring with the computer programs and print-outs purchased or licensed from the publishers. The student assessments which the teachers receive would include separate scores for vocabulary, reading comprehension, spelling, capitalization, punctuation, and language usage.

While such information manifested as digital line-scores for each student shown against coloured lines to indicate class or grade-level norms may seem to indicate a lot of hard, objective evidence of the student's attainment, it is important to remember the reality of the test situation which produces these results. When this test was reviewed by L. B. Birch of McGill University in *The Seventh Mental Measurement Yearbook* (Highland Park,

N.J., 1972), he noted that the "battery aims at the evaluation of generalized educational skills and abilities, not content achievement." He also commented that "the tests of capitalization and punctuation may soon be out of date," and warned of the danger that teachers might "fall into the error of teaching their classes along the lines of the test."

It should also be pointed out that like all objective, machine-scored tests, *CTBS* suffers from the necessity of having to test language items out of any meaningful context. Vocabulary scores, for example, are based on the student's matching of isolated words with their approximate synonyms, each one having several distractors. There are enough ambiguities in the choice of answers to make the reliability of the correct answers somewhat suspect. Similarly, spelling scores are based on the student's ability to recognize one standard spelling among four given 'non-words.' While this type of question may serve very well for the purpose of large-scale assessment, the question does not match the curriculum goal for language development which requires students to be able to spell the words that they use in their own written composition. Neither does this test show the ability of students to generate their own spellings for new words, using their knowledge of English spelling patterns and being aware that they are able to make corrections during the editing and proof-reading stages of the writing process. Teachers should also consider whether a spelling score based on the student's recognition of the forty-eight words chosen by the test-makers is a more valid indicator of the student's ability than the student's performance on classroom tests of the words actually used and studied throughout the curriculum.

CTBS, a Canadianized version of the *Iowa Tests of Basic Skills*, dates only from 1971. Nevertheless, *CTBS*, like many other objective tests of language, shows a somewhat out-dated view of language usage. Students are required to recognize whether words are used according "to the standards of correctly written English," but there is no accommodation in the test to a more modern view that language usage depends on the situational context. It completely ignores the question of what language level is appropriate for specific situations and the varieties of language function. This is the type of stress on "correct" usage which might demand an artificial classroom standard for "It is I," when most of the English-speaking world use, and most authorities accept, "It is me." The test demands that students distinguish between "me and him" and "him and me," but then includes an item in which "Andy and me" would be required in strictly formal English, although "Andy and I" would be acceptable at most levels of usage.

Bernard Spolsky, reviewing the *Canadian Achievement Test* in *The Sixth Mental Measurement Yearbook* (Highland Park, N.J., 1965), noted that "its blurred conception of usage will vitiate that test for anyone who is aware of current scholarship and attitudes in this field," and his comment could also apply to this section of the *CTBS*.

The comments made by respondents to the Survey suggest that some teachers appreciate having a battery of tests to administer each year under fairly controlled conditions across a number of classrooms, schools, or school districts. One respondent commented that it would be a "good standard" if "all students in the province do it," and "if done yearly, a measure of individual progress." Of course, it would be expected that any

student's scores would increase from year to year as part of the normal process of maturation, but in view of the general nature of the test items, it would not be possible to ascribe such growth to mastery of the language curriculum. Another respondent noted that the school uses "only the Vocabulary and Reading scores which are averaged to produce a composite grade level score for each child in June every year." Another respondent also noted that the school uses the reading and mathematics scores to help determine expectations for each grade and standards for "letter grade reporting."

The Survey shows that eighty-two per cent of the reported uses of *CTBS* occupied more than two hours of class time, or as one respondent commented, "an inordinate amount of class time." Another respondent also noted that the tests "are excessively lengthy for the amount of information which they provide." The most extensive comment was stated as the consensus of one school staff's views on the effect of external testing in general, and the use of *CTBS* in particular:

1. A loss of children's learning time.
2. A waste of valuable teacher time, marking and administering the test.
3. Presents an unfamiliar format to children--therefore the results are not meaningful.
4. We find primary children become anxious and frustrated with the format, length, and number of tests. Regular teacher-made tests give more accurate indication of the child's progress.
5. Diagnostic testing in June is impractical.

THE CANADIAN ACHIEVEMENT TESTS

The Canadian Achievement Tests (CAT), with 68 reported uses, presents many of the same issues as *CTBS* in its testing of reading comprehension, language skills, and mathematics, taking more than two hours of class time to administer, with the answer sheets being machine-scored and the results delivered by computer print-out of the Student's Test Record, the Class Record Sheet, the Objectives Competency Report, and the Group Response Record, all of which give the raw scores for each of the test objectives expressed as Grade Equivalent, National or Local Percentiles or Stanine rankings, or as a Scale Score. Responses to the Survey show that this test was regarded as compulsory by ninety-five per cent of the respondents, the Tests having been chosen mainly by school or district administrators (85 per cent), although in at least one district teachers were part of a district testing committee which was responsible for choosing the Tests. However, another respondent noted that the Test was decreed for all grade 10 students although the "district English teachers voted *against* having this test." The *CAT* battery was used most commonly in elementary schools (eighty-six per cent) although the eight levels (Levels 12-19) are designed to span grades 1.6 to 12.9.

The objections to the choice and use of this test are voiced in this respondent's comments:

"There is a very strong feeling against this particular test, as (1) May is a hectic month with many interruptions: band trips, assemblies, etc., and two or three classes lost to a test we don't use is resented. The results go to the Senior High. (2) Our grade 10's were all tested by the Reading assessment test last fall, so there is duplication."

Although there were other reports of *CAT* being used at the junior high level, most of the instances involved elementary schools. In some district-wide testing programs *CAT* was used for all grade 5 and 7 students, or for grades 3 and 6, but seven schools made it compulsory for all students in grades 4-7, five for grades 2-7, four for grades 2-5, two for grades 4-10, seven for grades 3-7, two for grades 1-7 and 2-9, and other schools for such combinations as grades 2-3, or grades 3-10.

As with the reports of the use of *CTBS* there is also considerable variety in the reported purposes for the use of *CAT*. Only about twenty per cent of the respondents perceived the test as being useful for program assessment. Others (twenty-four per cent) regarded the test as being aimed at the assessment of schools and classes, but the biggest percentage (thirty-two) saw the test as an assessment of individual pupils, with the results being used for either diagnosis or placement, hence the complaint of the respondent about the results coming too late in the year to be useful in the remediation of their students. One junior high school used the test only for the placement of grade 8-10 students in "special programs below the academic level" as in Modified English or remedial classes, while another elementary school used the test for pupil diagnosis and placement, and "to identify Native Indian students."

There is no doubt that the *Canadian Achievement Tests*, with all of the resources of modern computer programs, can provide teachers and administrators with a lot of information about the abilities of their students in comparison with the national or local norms established by the standardization procedures. Each of the test batteries is designed to measure achievement in reading, spelling, language, and mathematics, with a section on reference skills added for Levels 14-19. The *CAT* Technical Bulletin describes how the tests were normed in 1981 on a stratified random sample of schools across Canada, although it should be noted that in the breakdown of socio-economic characteristics of the student population, the overall percentage of students who speak a language other than English outside of school or come from homes in which a language other than English is spoken most of the time is only 7.5 per cent.

As with other standardized language tests, the test items, developed from the item pool of the *California Achievement Tests* on which *CAT* is modelled, are based on the objectives stated in "recent curriculum guides, textbooks, and other instructional materials used by school systems in all parts of Canada" (*CAT* Test Coordinator's Handbook). As an "achievement test" we should expect that the test measures the students' success in learning the knowledge and skills taught in a specific content area. Yet, of

course, the language items covered in a broad-based nationally standardized test might be quite removed from those actually studied in class. It is possible that a student might be very successful in learning to use the punctuation or usage conventions taught by the teacher, while the examples that appear in the test might be quite unfamiliar. Students could put a lot of effort into learning words in the class spelling program, or in spelling the words used in their own written expression, but still fail to recognize some of the twenty mis-spelled words on which CAT bases the scores for the students' spelling ability.

When the CAT Class Management Guide suggests that the "test scores represent achievement in basic skill areas," we must recognize that the material covered in these tests might be far removed from what the teacher considers to be "basic" in the student's development of oral and written language ability. The manual suggests, of course, that test scores must be reviewed together with the "student's actual classroom work," yet there is the danger that the computerized statistical results, because of their high-tech gloss, might take precedence, especially as one respondent noted, the CAT management program provides teachers with a copy of the student's scores printed on a gummed strip of paper to be stuck onto the student's permanent record file. When this happens, we can be sure that all cautions about the validity of the scores in relation to the whole language curriculum will be forgotten. Moreover, when statistical scores take on their own reality as concrete, scientific evidence of a student's achievement, then the test-makers' warnings about the reliability of these scores will also be ignored.

The CAT Technical Bulletin explains how the Kuder-Richardson formula 20 was used to determine the internal consistency of the test and to calculate the standard error of measurement. Tables 52-62 show the mean average scores, the standard deviations, and the KR 20 calculation for each of the tests in the battery across all of the eight levels of difficulty. In addition, the sample Student's Test Record shows each of the student's test results located in the centre of a band which represents the possible range of ability according to the Standard Error of Measurement. The teacher, in interpreting these results to the student or parent, is supposed to qualify any conclusions with the proviso that the built-in possibility of error in the student's results means that the student's "true" score might be as much as one grade above or below the reported Grade Equivalent score. Statistically, with tests that are standardized on the assumption that students' scores will follow the curve of normal distribution (an assumption which is itself questionable), the reported score may be within one standard error of measurement of range of the true score only two times out of three. The Ministry of Education's handbook *Selection and Use of Standardized Tests* also cautions teachers that there is only a 2:1 chance that the student's true score falls within one SEM (standard error of measurement) of the reported score. Would we have the same level of confidence in telling the time of day, as we have in labelling a student's language ability, if we knew that our watch, whether digital or analogue, had only a sixty-six per cent chance of giving the correct time, and then only to the nearest half-hour?

While such approximations of true scores are quite acceptable in the assessment and comparison of large groups of students, consideration of the

CAT's standard errors of measurement will mean that a teacher should hedge any conclusion about an individual student's scores with the knowledge that there is only a two-thirds chance that the score will be within a half or more of a grade level of what the student is really capable. With such qualified acceptance of any score, it is easy to see why assessments of the student's in-class work and test results may make more reliable indicators of the student's achievements.

Teachers also need to be aware of the content validity of these tests. While the CAT Test Coordinator's Handbook asserts that "test items should be answered as readily by students taught by a traditional approach as those who were taught by any of the newer approaches," the teacher needs to assess to what extent the test items actually reflect the content of the curriculum. In fact, CAT recommends that the examiner should write the test before administering it to the students. CAT also recommends the use of Locator Tests (as were used during the process of standardizing the national norms) to ensure that students are given the right level of test battery for their ability. Test results for students who perform at either the low or high end of the scale at any particular level become extremely unreliable. In some of the tests the difference of one correct answer may make a difference of one and a half grades in the student's grade equivalent score.

The meticulous procedures by which the *Canadian Achievement Tests* are prepared, pre-tested, vetted for racial, ethnic or sex bias, and standardized with an item-by-item analysis of results after the tests had been piloted in 171 schools across all of the regions of Canada have resulted in an impressive testing program. In many ways these tests represent a great advance over previous such test batteries, especially in the selection of reading passages for the comprehension tests. The amount of statistical information that can be gathered for every student, class, or school in a district is equally impressive. Details of each student's scores on each of the sub-tests become immediately accessible. So, too, does a class's performance in meeting CAT's criteria for competence or proficiency in each of the tested skills become available to the teacher. Also very useful are the suggestions contained in the Class Management Guide for classroom activities by which the teacher can help students to develop their skills in specific categories spotlighted by the test results.

Because the Canadian Achievement Tests can provide such useful data for the teacher or administrator, it is extremely important that teachers realize the limitations of such a test. There is so much that the test can do, but within the context of a whole language program, there is still a great deal that it cannot do. It is essential to realize that these tests measure only the students' responses to the examiners' choice of language; they do not measure directly the students' abilities to generate their own language.

Nowhere are the safeguards against misuse and abuse of test data more necessary than in the recording, reporting, and interpretation of scores. The Canadian Council of Teachers of English makes the same point in its *Evaluation Policy* (1985):

For all standardized tests or other external tests employed, each teacher, principal or counsellor making use of test data must be

acquainted with the basis on which norms were developed and the error of measurement provided in the technical manual for the test, and ensure that scores are reported, used, explained (to parent or student) in a manner that makes clear the limits of precision.

In such a battery of tests as *CTBS* or *CAT*, there are some differences in the relevance of particular tests. "Reading," for example, whether it means word recognition, vocabulary, or the comprehension of short passages, is adequately labelled for the specific skills it measures, and the teacher or parent should be aware of other dimensions of the reading program which are not included in the test. "Language," however, is divided between mechanics and expression, and shares the limitations of many earlier tests of English language reviewed by J. N. Hook in *Reviews of Selected Published Tests in English* (N.C.T.E., 1976), who made this comment about the nature of such tests:

Very noticeable is the narrow coverage of the tests. One finds nothing about dialects. Nothing about the history of the language. Almost nothing directly about semantics, unless one counts vocabulary items as semantics. Nothing about etymology. Little, except in a very few tests, about the actual working of the English sentence.

The very nature of standardized tests designed for assessing large numbers of students demands that students respond by choosing one answer among several distractors, or marking the correct item in a given sentence. Students' answers must be right or wrong. There is no middle ground, and yet teachers of language know that many questions of language use do not yield to such clear-cut arbitration as in mathematics. Hook questions the view of language that is pervasive in objective tests:

The items in these tests ask the students to determine whether or not something is "correct": a spelling, a choice of verb, an arrangement of sentence parts, etc. These tests confirm the stereotype of the English teacher as someone primarily interested in catching someone making errors: "Oh, if you're an English teacher, I'd better be careful of what I say."

It must also be remembered that all of these language tests, whatever their level of technical sophistication, measure only a part of the students' dealings with written language. The results of the Survey clearly indicate the total neglect of oral language, despite all that is known about the processes by which children acquire and develop their language abilities through oral interaction. Walter Loban, in "Language Development and its Evaluation," points out that "the living language is the spoken language," and warns of the danger of letting the focus of standardized tests deflect our teaching and curricula from the central concerns of language development by giving undue importance to minor objectives:

To reduce power over language to such mechanics of written language as spelling, punctuation, and capitalization is a dangerous oversimplification. These are *not* the true fundamentals of language. The people of this nation need instruction that focuses not only upon details but also upon larger adaptations, such as vigor of thought and precise expression of thought and feeling. A perspective that begins with errors of mechanics rather than with a more complete picture of desirable accomplishment seldom reaches to the really important aspects of language ability--interest, pleasure in doing or using, organization, purpose, and other crucial integrating and dynamic patterns of performance.

READING TESTS

PROVINCIAL LEARNING ASSESSMENT PROGRAM (PLAP)

Although the Survey was designed to cover tests administered during the 1984/85 school year, there were 57 responses referring to the 3. C. Reading Assessment which was carried out in May, 1984, with the Summary Report becoming available in January, 1985. Nevertheless, it is worth considering the comments on this assessment program because the test materials remain with the school districts and it is possible that a district might decide to re-test its students at these grade 4, 7 and 10 levels.

One school noted that ESL students were not given the test, and another respondent noted that mentally handicapped and retarded children were excluded from taking the test. In districts where such students are mainstreamed into regular classrooms it may be left to the individual teacher's judgment as to which students actually take the test, and which are excused on humanitarian grounds.

A positive comment was that these tests were "reasonable in length and easy to administer," enabling school principals to compare their schools' performances against district and provincial averages, and noting that "the reading tests particularly may point to some weaknesses in the basal reader approach this district uses." Another respondent commented that "the ministry tests have been useful as overall additional evaluation of programs/students of the school."

In addition to the assessment of students' abilities in understanding words, their literal, inferential, and critical comprehension, and their study and research skills, the 1984 Assessment also repeated the use of the Estes Reading Attitude Scale to measure students' attitudes towards reading, as was done in the 1980 Assessment. One important innovation of the 1984 Assessment was the use of open-ended responses for a random sample of four per cent of the 98,000 students involved in the test. While the marking of such a test instrument is much more complicated than the machine-scoring possible for multiple-choice questions, nevertheless the

open-ended format is necessary for the measurement of students' abilities for many important objectives of the reading program, especially "whenever student ability to respond, to substantiate, to organize, or to integrate is of interest" (*Summary Report*).

The authors of the *Summary Report* also point out that the use only of fixed-response questions may create wrong impressions about what is important in the curriculum:

While the multiple-choice format is the single most effective and widely used technique for wide-scale assessment, it prevents consideration of many important objectives, and causes others to be measured in artificial and unnatural ways. Perhaps more importantly, it presents the implicit message that the Ministry and associated agencies believe that objectives which are amenable to fixed-response are the most--or, indeed, the *only*--important objectives, or, alternately, that it is the only method which permits reliable and valid measurement. Both beliefs are mistaken; many important objectives are best measured using open-ended response formats; valid and reliable procedures are available.

CRITERION-REFERENCED READING TESTS

Fifteen respondents reported using the Ginn Mastery tests for students involved in the Ginn 720 Reading Program. Most of these responses represent decisions made by the teacher concerned, perhaps on the recommendation of district personnel. Typically this type of test is used by the classroom teacher to ensure that a student has achieved mastery at a specified level of the reading program before going on to a more difficult reading textbook. In some cases the test might be used to diagnose a student's reading ability in order to determine the proper entry level into the reading program. Five schools reported using the mastery tests for all students, grades 1-7, in the school. In these cases the test was chosen by district administrators, although teachers administered and marked the test. The results were used by teachers for pupil assessment, but at the same time the scores were collected at the district level, presumably as a means of comparing school or class performance.

Considering that a large proportion of the 1165 elementary schools in the Province use the Ginn 720 Reading Program, it is surprising that there were not many more reports of the use of the Mastery Tests which are an integral part of the program. In a typical elementary classroom in which the range of the students' reading abilities spans many levels of the program, the use of mastery tests enables individual students to progress from one level to the next according to the development of their reading skills. It is possible, however, that because the test is such a normal part of classroom routine, most teachers did not regard it as an "external" test for this Survey.

STANDARDIZED READING TESTS

The list of reported uses of commercial tests (Appendix B) shows the great variety of reading tests that were used in British Columbia during 1984-85. Some were chosen by teachers to assess the level of reading ability of the students in their classes. Others were used by learning assistance teachers to assess and diagnose those students who might need remedial help. In some cases, acting on a decision by a school staff, all of the students in the school would be tested at the appropriate level of a given test. In other cases a standardized reading test was chosen by district personnel to be administered to all of the students at specified levels in the district.

The most widespread test reported in the Survey is the Gates-MacGinitie Reading Tests, with 199 reported uses. These tests were first developed at Columbia University in 1926, revised and restandardized in 1965, with a machine-scored edition in 1971 necessitating a new set of norms. A Canadian version was developed in 1979. The eight levels of these tests span all of the grades from reading readiness at the beginning of grade 1 to the Survey F for grades 10-12. All levels are tested for Vocabulary and Comprehension, while the intermediate and secondary levels also have tests for Speed and Accuracy.

In B.C. these tests were used mainly at the elementary level (85.9 per cent), with only nine reports of the tests being used for secondary schools and two reports of its use for K-12. In 70.4 per cent of the cases the tests were compulsory for all students at the specified grade level. In 26.6 per cent of the responses the choice of the test was made by teachers, the remainder being chosen by school or district administrators, or by a combination of both. Marking of the tests was done mainly by teachers (78.4 per cent), with 4.5 per cent being marked by markers, and only 17.1 per cent by machine-scoring. The purpose of the test program was given mainly as pupil assessment (71.4 per cent), but class and school assessment were also seen as the purpose in 50.8 per cent of the responses.

Nevertheless, one respondent commented: "We rely on teacher tests that test our own objectives. This test we realize gives us very limited information." Another report of these tests being used at the junior high level included the comment: "We find the results of the Gates-MacGinitie of little value," although an elementary school reported using the tests for their grade 7 students because it was "requested by the Secondary School receiving these students. . . . to be used to help placing students in specific programs and help identify students for remediation, etc."

Like most of the other reading tests that have shown up in the Survey, and like the reading sections of the *CTBS*, *CAT*, and Metropolitan test batteries, the Gates-MacGinitie Tests project a very specific view of what constitutes reading ability, and its measurement of these skills shares the same drawbacks and limitations as were noted in the comments on *CTBS* and *CAT*. That the authors of the Teacher's Manual were aware of the tests' limitations is shown in their caution regarding the interpretation of students' scores:

Several cautions regarding the interpretation and use of grade scores must be given. First of all, the grade scores given in the norms are simply average values; they do not indicate how well a given student or class *should* read.

The authors are also very cautious in warning teachers about the possible errors in measurement that can affect the reliability of the test scores:

Several types of errors influence a student's score on a test. The items included in a test constitute only a small sample of all possible items. No matter how carefully the items were written and selected, a different sample of items could well result in somewhat different scores for some students. Also variations in a student's performance are a cause of unreliability in test scores. A student's score may be influenced by his physical condition, by a distraction occurring during the testing, and by many other factors. In general, all test scores, and therefore all interpretations, should be regarded as tentative.

Unfortunately, when a student's score is printed out by a computer for attachment to the student's cumulative record card, such provisos about the tentativeness of the scores may well be forgotten.

Many of the other commercial reading tests that were reported in the Survey show the same characteristics as the Gates-MacGinitie, and indeed, the reading sections of the comprehensive test batteries. Generally such tests measure the recognition of words, or at the primary levels the identification of the sounds of letters at the beginning or ends of words. Since the Gates-MacGinitie Tests were developed some sixty years ago there has been an attempt to introduce more variety or relevance into the reading passages on which the students are tested for their comprehension. Some reading tests include literary passages or short stories; others may use short poems, usually in the lyrical or narrative mode. But with these items, of course, it is impossible within the format of an objective, machine-scored test to measure the higher levels of comprehension beyond the literal or inferential categories of response to reading.

An interesting variation which may deserve more attention is the use of *GAPADOL* or the GAP Reading Comprehension tests by three respondents. In these tests, instead of the student being required to choose the given correct answer from amongst several detractors, the student is required to supply a word which will fit the sense and syntax of a space in a given sentence or passage. This technique for measuring reading is known as a Cloze Test, and is commonly used by teachers as a means both of teaching and testing of students' abilities to handle the level of difficulty in syntax or vocabulary of a specific literary text. The technique is also used in conjunction with other subjects across the curriculum when students are taught how to tackle the content, for example, of science or social studies texts.

Because the student has to supply the word or phrase to fill the gap in the text, *GAPADOL* may present some technical problems in the marking or standardization of students' scores. Nevertheless, the loss of statistical efficiency may well be compensated by the fact that this type of test more closely mirrors the contemporary view of the reading process as a creative, generating activity on the part of the reader. It is also of some interest that although this is an Australian test, the norms were obtained through standardization in Saskatchewan.

Apart from the Gates-MacGinitie Tests, the Survey showed that most of the commercial reading tests were chosen by teachers, often in conjunction with school or district administrators. Because of the wide choice of commercial tests available on the market, teachers, especially those in such specialized areas as ESL or ABE, are able to choose materials that fit their particular purposes in the evaluation or diagnosis of their students' language abilities, and to use the results with some degree of circumspection, because any scores obtained within the classroom use of such tests can be confirmed by the teacher's direct observations of the student's language.

SUMMARY

The Survey of the Testing of English/Language Arts in B. C. during the 1984-85 school year shows a wide variety in the use of tests. To the extent that the response represents a typical sample of schools and colleges throughout B.C. it can be seen that a considerable proportion (9.1 per cent) manage to avoid all externally-imposed tests, while another 9.8 per cent reported using only the tests mandated by the Ministry of Education. The remainder is divided between tests imposed either by the district or school administrators. In both cases the teacher may be required to administer either tests developed by the administrators for a specific purpose, or they could be required to use a commercial test bought off the publisher's shelf.

What is tested in the locally-developed tests is not revealed in this Survey. The curricular content validity of district-wide tests can only be evaluated by an intensive analysis of the test items in relation to the goals and objectives of the programs being evaluated. Such an analysis of both the test instruments and the local curricula is beyond the scope of this Survey. However, it is important to note how widespread is the use of district tests, and to consider the purposes served by the test data.

Because the Survey revealed the widespread use of such tests as *CTBS*, *CAT*, *The Metropolitan Achievement Test*, and the *Gates-MacGinitie Reading Test*, it was possible to comment on both the content validity and the reliability of these tests. Where such standardized tests are used for large-scale assessment of large numbers of pupils, any discrepancies caused by the individual difficulties of pupils will, in the long run, be evened out in the statistical procedures for calculating class, school, or grade norms. That compilers of standardized tests deliberately choose test items to create the largest discrimination amongst the pupils does not matter. Only when the fate of the individual student in the classroom is at stake does the fact that

a "good" test item is one on which fifty per cent of the target population is doomed to fail become a significant issue.

A classroom teacher, on the other hand, may be very happy to find that after teaching a particular concept, the great majority of the students in the class are able to answer correctly an item designed to test that concept. But for the compiler of a standardized test, such an item might be thrown out of a test because it does not separate the sheep from the goats.

It is also apparent from a cursory reading of technical bulletins, that the standard error of measurement in all standardized tests must be kept to the fore in all reporting and recording of individual students' marks. The possibility of error because of the built-in unreliability of standardized tests when applied to the single case is an ever-present danger. When such scores are used to give the students' report-card grades, or worse, when these scores are used to slot students into ability groupings within a class or a school, there is always the danger that these students will live up (or more likely, down) to the mistaken expectations of a test score.

It is also necessary to examine more closely the match between test and curriculum. It is patently unfair to students to teach according to one set of curricular objectives and to test according to another set, or sub-set. This is particularly true of tests of English curricula in which oral objectives are an important part of the course. It is disappointing that there were no responses to the Survey that showed that any level of school, district, or provincial administration had paid any attention to the assessment of oral skills. Obviously, B.C. administrators might learn from the Manitoba provincial assessment program for the evaluation of students' oral abilities, or the work of the Language Monitoring Team of the Assessment of Performance Unit in the U.K.

The assessment of writing skills does show some promise, especially in those district programs in which samples of students' writing in various modes of discourse are collected for evaluation and analysis. On the other hand, the Survey also shows that there are some administrations which still believe that writing ability can be assessed through objective, multiple-choice tests. It is a pity that the news about student writing folders or folios, as mandated in Ontario and used in national CSE examination in the U.K. and by the Educational Testing Service in the U. S., has not spread throughout B. C. While the value of this assessment technique at the elementary and secondary levels is well known, Robert Camp, of the ETS, suggests that the writing folder may well have important applications in post-secondary assessment. In her report to the CCTE National Seminar on Evaluation she pointed out it could meet the ETS objectives to provide a "broader range of information about students' talents," "to emphasize writing in the secondary/post secondary transition," and to "provide a more comprehensive and more reliable source of information for admissions than is currently provided by applications essays." (See *Directions and Misdirections in English Education*, ed. Peter J. A. Evans)

The biggest weakness in the Survey is its failure to show clearly the real purpose behind each of the tests reported in the questionnaire. Perhaps it was naive to assume that the respondents to the Survey would have a

clear understanding of the primary purpose behind the administration of an external test. Or it may have been an optimistic assumption that a survey of the world of mass testing and its heavy reliance on multiple-choice questions would necessarily encourage respondents to choose the single best answer to the questions about a test's purpose rather than a scatter of possible uses among the available choices. Or, as it has already been suggested, it may be likely that there is a genuine confusion about the purpose of testing programs. It may well be that because a test *could* be used for a variety of different purposes, then the respondents assume that the test *is* being used for all of these purposes. In any case the percentages given in Table 4 must be used only with caution, because they are derived from combining the responses on questionnaires in which more than one choice was indicated.

The question of purpose is the one area most worthy of some follow up research, because all of the other questions about the validity and reliability of a test instrument are to some extent dependent on to what purpose the test data are put. The question of purpose needs to be asked specifically about every test instrument used as the basis for examining the suitability of that instrument for that purpose. This is especially true of locally-developed school or district tests which are not normally available to the independent observer. Each of these tests needs to be analyzed according to its stated purpose. Every test item should be thoroughly examined to see if it contributes to the achievement of that purpose.

QUESTIONS FOR FURTHER CONSIDERATION

Because the Survey was focussed on the specific test instruments which are used to assess students' skills in language, there remain many questions about the distribution of the schools and school districts which use these instruments, the numbers of students affected by their testing programs, and the total cost of in terms of materials, computer time, and teachers' time needed to implement such programs.

It must be emphasized that this Survey is only a beginning, and while it provides some useful information about the use of various types of external tests across the province, it still leaves many questions unanswered. Perhaps the biggest puzzle about mass testing is what happens to the test data gained at considerable cost in money and time? The provincial assessment program is aimed at the improvement in curriculum and instructional methods, and it is possible to point to the boxes of instructional materials that have been distributed to school following the assessments of Written Expression and Reading. Whether these have had much impact on curriculum is difficult to say, although the Writing Evaluation Folio Kit for use in student placement with the new Grade 11 writing course does have direct applications, and follows the lessons learned through The Young Writers Project. District assessments are also ostensibly dedicated to program assessment, but it remains to be shown just what impact the assessment programs have upon curriculum and instructional methods. Are the test data gathered, distributed, and then filed away? Or are the test data useful in the allocation of funds or resources? How are the findings of a district assessment program translated into the improvement of educational practices and materials?

Obviously, more research will be necessary to complete the picture of the state of language testing in B.C. And, of course, a snap-shot taken of the kaleidoscope of testing gives only an impression in one time-frame, because assessment programs and test instruments are themselves in a state of flux, and subject to change. It is all the more important, therefore, that teachers, administrators, students, parents, trustees, and members of the public should become aware of the issues involved in the assessment of language ability, and consider the following questions:

1. How much external testing of English/Language Arts is necessary to augment the evaluation procedures which elementary and secondary teachers use as part of their instructional methods?
2. Are the test data currently being collected by English/Language Arts tests used to revise curriculum, develop instructional materials, or improve teaching methods?
3. Should methods and materials be developed for assessing students' abilities in oral language?
4. What are the costs, both financial and human (in terms of the teachers' and students' instructional time) of a district's testing

program for English/Language Arts?

5. Should a testing program for English also include an assessment of the students' abilities to understand, criticize, and evaluate the use of language in print, film, and electronic media?

REFERENCES

- Buros, Oscar K., ed. *The Sixth Mental Measurement Yearbook*. Highland Park, N.J.: Gryphon, 1965.
- Buros, Oscar K., ed. *The Seventh Mental Measurement Yearbook*. Highland Park, N.J.: Gryphon, 1977.
- Canadian Council of Teachers of English. *Evaluation Policy*. 1985.
- Cooper, Charles. Measuring Growth in Writing. *English Journal*, 64: 111-120 (1975).
- Cooper, Charles R. and Odell, Lee. *Evaluating Writing*. Urbana, Illinois: N.C.T.E., 1977.
- Diederich, Paul B. *Measuring Growth in English*. Urbana, Illinois: N.C.T.E., 1974.
- Evans, Peter J. A., ed. *Directions and Misdirections in English Evaluation*. C.C.T.E.: Monographs and Special Publications, 1985.
- Grommon, Alfred H., ed. *Reviews of Selected Published Tests in English*. Urbana, Illinois: N.C.T.E. 1976.
- Hook, J. N. Tests on the English Language. *Reviews of Selected Published Tests in English*. ed. Alfred H. Grommon. Urbana, Illinois: N.C.T.E., 1976.
- Loban, Walter. Language Development and Its Evaluation. *Reviews of Selected Published Tests in English*. ed. Alfred H. Grommon. Urbana, Illinois: N.C.T.E., 1976.
- Ministry of Education. *Selection and Use of Standardized Tests*. Victoria: Province of British Columbia, 1979.
- Ministry of Education. *Grading Practices: Issues and Alternatives*. Victoria: Province of British Columbia, 1979.
- N.C.T.E. *Common Sense and Testing in English*. Report of the Task Force on Measurement and Evaluation in the Study of English. Urbana, Illinois: N.C.T.E., 1975.
- Stibbs, Andrew. *Assessing Children's Language*. London: Ward Lock Educational, 1979.

Please complete a copy of this form for each external (i.e. originating from outside the classroom) test of English/Language Arts, including reading, used during the school year 1984/85.

THE COMMISSION ON TESTING OF ENGLISH/LANGUAGE ARTS

Survey of External Tests Used in B.C. 1984/85

A. Name of the Test:

B. Source of the Test:

C. Student Population Involved in this Test:

Please give the approximate number of students involved, and their categories or levels.

Please check as many boxes as apply to this Test.

1. Is this Test compulsory?

1. Yes ☐ 2. No ☐ 3. For all students? ☐

4. For specific groups of students ☐ Please identify: _____

2. What are the purposes of this Test:

1. Program Assessment ☐ 2. Class Assessment ☐ 3. School Assessment ☐ 4. Pupil Assessment ☐

5. Pupil Diagnosis ☐ 6. Pupil Placement ☐ 7. Other ☐ Please specify: _____

3. How was this Test chosen:

1. By teachers ☐ 2. By school administrators ☐ 3. By district administrators ☐ 4. By the Ministry of Education ☐

5. Other ☐ Please specify: _____

4. How is the Test marked?

1. By teachers ☐ 2. By markers ☐ 3. By machine-scoring ☐

4. Other: ☐ Please specify: _____

5. How are the results recorded?

1. For individual students ☐ 2. By classes ☐ 3. By schools ☐ 4. By school districts ☐

5. Other ☐ Please specify: _____

6. How are the results distributed?

1. To students ☐ 2. To parents ☐ 3. To teachers ☐ 4. To school trustees ☐ 5. To other institutions ☐

6. To school counsellors ☐ 7. To district staff ☐ 8. To school administrators ☐

9. Other ☐ Please specify: _____

7. Who has access to the results?

1. Students ☐ 2. Parents ☐ 3. Teachers ☐ 4. Administrators ☐ 5. School trustees ☐

6. Other ☐ Please specify: _____

8. How much class time is spent in taking the Test?

1. None ☐ 2. Less than 1 hour ☐ 3. 1-2 hours ☐ 4. More than 2 hours ☐

Do you wish to receive a copy of the Commission's Report? _____

Name and Address of the School or College: _____

Please add overleaf any comments or further information about the testing of English/Language Arts. The Commission will be pleased to receive your opinions about this Survey.

Would you be willing to be interviewed by a member of the Commission to discuss the impact of this Test on your school's programs?

If "Yes" please give your name and telephone number:

Note: Individual respondents will not be identified in the Commission's report).

Please return this form to the B.C.T.F.

APPENDIX B

REPORTED USES OF COMMERCIAL TESTS

<i>Adult Basic Reading Inventory</i>	2
<i>Barnell-Loft Spelling Test</i>	1
<i>Bilingual Syntax Measures I and II</i>	1
<i>Boehm Test of Basic Concepts</i>	2
<i>Botel Reading Test</i>	1
<i>Brigrance Diagnostic Inventory of Basic Skills</i>	5
<i>B.C.Quiet</i>	2
<i>Canadian Achievement Test</i>	70
<i>Canadian Cognitive Abilities Test</i>	2
<i>Canadian Test of Basic Skills</i>	264
<i>Criterion Test of Basic Skills</i>	1
<i>Doren Diagnostic Reading Test of Word Recognition</i>	1
<i>Developmental Review Program</i>	2
<i>Diagnostic Achievement Test in Spelling</i>	1
<i>Diagnostic Reading Tests</i>	2
<i>Durrell Analysis of Reading Difficulty</i>	3
<i>Dolch List of 220 Basic Sight Words</i>	1
<i>GAP Reading Comprehension</i>	3
<i>Gates-MacGinitie Reading Test</i>	199
<i>Henmon-Nelson Reading Test</i>	1
<i>Industrial Reading Test</i>	1
<i>Illinois Test of Psycholinguistic Abilities</i>	1
<i>Jackson Aptitude Multi-Dimensional Battery</i>	1
<i>John's Basic Reading Inventory</i>	1
<i>Metropolitan Achievement Test</i>	7
<i>Metropolitan Reading Readiness</i>	6
<i>Nelson-Denny Reading Test</i>	29
<i>Otis Lennon School Abilities Test</i>	1
<i>Peabody Individual Achievement Test</i>	8
<i>Peabody Picture Vocabulary Test</i>	3
<i>Roswell Chall Diagnostic Reading Test</i>	1
<i>Reading for Understanding Placement Test</i>	3
<i>Schonell Graded Spelling Test</i>	2
<i>Slosson Oral Reading Test</i>	2
<i>Slingerland Screening Test</i>	2
<i>Stanford Achievement Test</i>	23
<i>Stanford Diagnostic Reading Test</i>	48
<i>Test of Written Language</i>	3
<i>Test of Reading Comprehension</i>	1
<i>Van Wagenen Reading Test</i>	1
<i>Wepman Auditory Discrimination Test</i>	1
<i>Watson Glaser Critical Thinking Appraisal</i>	1
<i>Watson Reading Readiness Test</i>	2

APPENDIX C

Table 1 Types of Tests Reported

<i>Provincial</i>	88	9.1%
<i>District</i>	39	4.0%
<i>Reading Mastery</i>	15	.15%
<i>School</i>	12	.12%
<i>Commercial</i>	712	74.7%
<i>EPT</i>	13	.13%

APPENDIX D

Table 2

Implementation of Commercial Tests

	<u>Compulsory</u>	<u>Voluntary</u>
<i>CTBS</i>	92%	8%
<i>CAT</i>	95.6%	4.4%
<i>Gates-MacGinitie</i>	70.4%	29.6%
<i>Stanford Achievement</i>	64.7%	35.3%
<i>Stanford Diagnostic</i>	76.6%	23.4%
<i>Nelson-Denny</i>	84.6%	15.4%

APPENDIX E

TABLE 3 - CHOICE OF TESTS

	CTBS	CAT	GATES-MACG
By Teachers	7.2%	4.4%	26.6%
By School Administrators	15.2%	5.9%	11.6%
By District Administrators	40.4%	32.4%	28.6%
By Teachers and AdministratorsS	33.5%	52.9%	29.1%
By Ministry	3.8%	2.9%	2.5%

APPENDIX F

Table 4 - Purpose of the Test

	CTBS n=263	CAT n=68	Gates-MacG n=199
Program Assessment	4.1%	0.9%	2.3%
Class Assessment	4.9%	10.9%	7.1%
School Assessment	18.2%	18.1%	8.4%
Pupil Assessment	32.9%	27.2%	36.7%
Pupil Diagnosis	27.4%	22.7%	22.3%
Pupil Placement	12.3%	20.1%	22.8%